Preparing, Restructuring, and Augmenting a French Treebank: Lexicalised Parsers or Coherent Treebanks?

Natalie Schluter
NCLT, School of Computing
Dublin City University
Dublin, Ireland
nschluter@computing.dcu.ie

Josef van Genabith
NCLT, School of Computing
Dublin City University
Dublin, Ireland
josef@computing.dcu.ie

Abstract

We present the Modified French Treebank (MFT), a completely revamped French Treebank, derived from the Paris 7 Treebank (P7T), which is cleaner, more coherent, has several transformed structures, and introduces new linguistic analyses. To determine the effect of these changes, we investigate how the MFT fares in statistical parsing. Probabilistic parsers trained on the MFT training set (currently 3800 trees) already perform better than their counterparts trained on five times the P7T data (18,548 trees), providing an extreme example of the importance of data quality over quantity in statistical parsing. Moreover, regression analysis on the learning curve of parsers trained on the MFT lead to the prediction that parsers trained on the full projected 18,548 tree MFT training set will far outscore their counterparts trained on the full P7T. These analyses also show how problematic data can lead to problematic conclusions–in particular, we find that lexicalisation in the probabilistic parsing of French is probably not as crucial as was once thought (Arun and Keller (2005)).

1 Introduction

The construction of the Paris 7 Treebank (P7T) brought to fruition the first treebank available for French (Abeillé et al. (2004); Abeillé and Barrier (2004)). Its use in research, however, has proven challenging. Arun and Keller (2005), for example, observe a number of points in which the treebank should be improved or even completely structurally reorganised before any serious study can be carried out using it.

Our goal has been to create a French treebank with consistent and coherent annotation and with a comparatively low error rate, that supports efficient statistical parsing paradigms while compromising as little as possible on linguistically relevant structural information. We hope to have done this, while carrying out only the minimum number of changes to the P7T necessary to meet this goal.

The necessary correction and modification of the P7T has led to the creation of the *Modified P7T*, which we will simply call Modified French Treebank (MFT). Our research focusses on the functionally annotated subset of 9357 sentences from the P7T, and the MFT now consists of the the first half of these sentences.

Following an overview of the P7T (Section 2), we introduce the MFT via the various structural changes (Section 3), formatting and error mining (Section 4). Using statistical analysis techniques, we show that the MFT and P7T have become very different treebanks (Section 5). Finally, as a means of showing the importance of such changes in treebank-based linguistic analysis, we give results for statistical parsing in Section 6, and draw some important conclusions.

2 The Paris 7 Treebank

Work on the P7T was carried out by a research team at the Université Paris 7, under the direction of Anne Abeillé. The treebank consists of *Le Monde* newspaper article excerpts published between 1989 and 1993, written by various authors, and covering an array of topics. The full P7T contains 20,648 sentences annotated for phrase structure, (and additionally, about half with grammatical function tags) comprising 580,945 words. Table 1 gives the phrase tags of the P7T. We notice, in particular, that there is no VP, except in the cases of some participial phrases (VPpart) and infinitival phrases (VPinf).[1]

[1] The phrase VN is considered to be more of a convention, grouping together all parts of composed verbs into one unit with their clitic pronouns, as well as any modifier phrases

Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics, pages 200–209.

AP	adjectival phrase
VPinf	infinitival phrase
AdP	adverbial phrase
Srel	relative clause
COORD	coordinated phrase
Ssub	subordinated clause
NP	noun phrase
Sint	internal, inflected sentence
PP	prepositional phrase
VN	verb kernel
VPpart	participial phrase
SENT	independent sentence

Table 1: Phrase Tags of the Paris 7 Treebank

Table 2 gives the syntactic function labels used in the functionally annotated sections of the P7T. Only some clitics and those phrases which are sisters of a VN constituent carry functional annotations. This assumes that any phrase which is a sister element of VN functionally depends directly on the verb kernel; we show that this is not always the case and present a new functional annotation scheme in Section 3.5.

SUJ	subject	DE-OBJ	de (of/from) -object
OBJ	object	A-OBJ	à (to)-object
P-OBJ	preposition-object	MOD	modifier
ATS	subject attribute	ATO	object attribute

Table 2: Syntactic Function Labels of the Paris 7 Treebank

Our project focusses on the first half of the functionally annotated sentences of the treebank; there are, in total, 20 files that contain the 9357 functionally annotated sentences, and we are working with the first ten of these files. These files originally contain 4741 sentences, comprising 134,445 words.

3 Structural Changes

The MFT differs significantly from the P7T, in terms of its phrase structure as shown by the statistical tests in Section 5. Major structural changes to the original P7T trees include increased rule stratification, introduction of analyses for untreated structures, information propagation, coordination raising, the addition of missing functional tags, and the introduction of functional path tags.

occurring between these.

3.1 Rule Stratification

While maintaining a relatively flat syntactic analysis, the MFT has the property that there is one distinct head (and sometimes also one co-head) for each constituent. For example, NP, AP, and AdP constituents that have modifiers will have separate constituents for those modifiers. Figure 1 provides an example of increased stratification for AdP in Example (1).

(1) encore pas très bien
still not very well
'still not very well'[2]

3.2 Introduction of Analyses for Untreated Structures

Compared to the P7T, the MFT offers increased coverage of linguistic phenomena. 'It'-cleft constructions are an example of structures that remained untreated in the P7T annotation guidelines, and therefore received a variety of treatments throughout the P7T. Figure 2 (for Example (2)) illustrates the new analysis, inspired mainly by van der Beek (2003).

(2) C'est [...] l'URSS [...] qui se
It is [...] the USSR [...] who herself
trouve prise [...]
finds taken [...]
'It is the USSR that finds itself trapped'[3]

3.3 Information Propagation

Some constituent categories in the P7T derive terminal strings with grammatical patterns not reflected in the intervening levels of syntactic representation. VPinf, VPpart, and Srel are the three categories which were found to have this property. For instance, VPinf necessitates a VN daughter that has a V daughter which is an infinitive, and Srel necessitates a PP or NP daughter whose head is or has an argument that has a relative pronoun daughter. Such cases amount to information loss across levels of representation, thereby introducing CFG ambiguity. A parser must guess the daughters of these constituents VN, NP, and PP, in order to carry out correct annotation. This potentially leads to poor statistical parsing. We automatically propagate the required information, aug-

[2] Sentence 88, file flmf7ag1ep.cat.xml.
[3] Sentence 8151, file flmf3_08000_08499ep.xd.cat.xml.

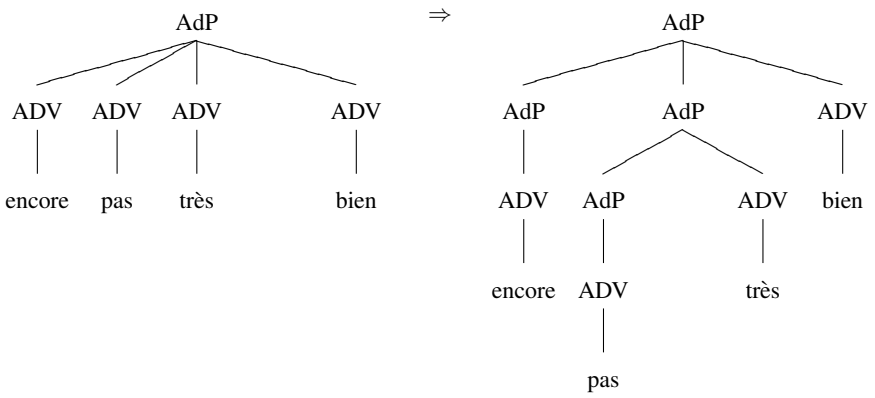

Figure 1: P7T representation (left) and MFT representation (right) of example (1).

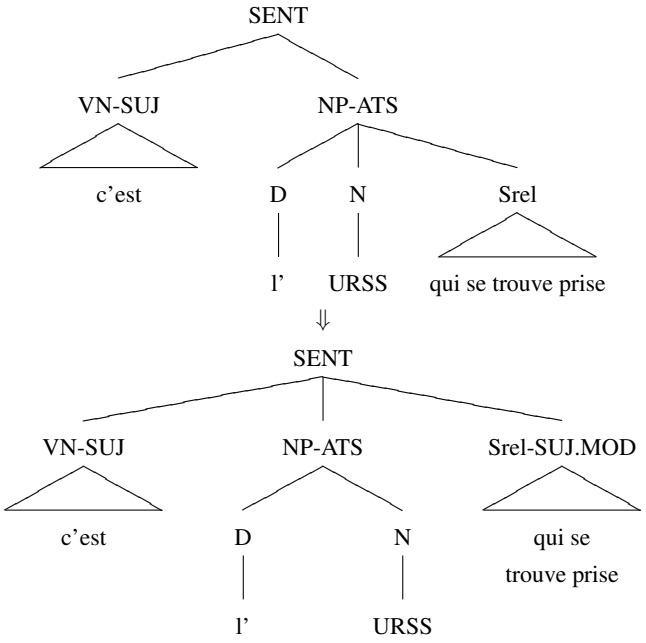

Figure 2: P7T representation (above) and MFT representation (below) of example (2).

menting the MFT with these constituent labels.[4]

- In the MFT, the part-of-speech V is separated into three different categories: Vfinite, Vinf, and Vpart, according to whether the verb is tensed, infinite, or a participial. The XML representation of the constituent VN for the MFT now has an attribute "type", which records which is the first verb POS: finite, inf, or part. Now the VPinf constituent will only have a VN constituent with type "inf", and similarly for VPpart.

- Relative pronouns in the P7T are already indicated in the "subcat" attribute. We propagate this information as "type" attributes through the dominant nodes, until the dominant node Srel is reached, thus introducing the constituent categories PPrel and NPrel.

Example (3), whose tree structure is shown in Figure (3), illustrates both these changes.

[4]Note that this is similar to the strategy suggested by Johnson (1998), but with two important differences. First, the information propagation is done here in a bottom-up fashion and, therefore, retains the central linguistic motivation behind phrase structure trees, that of constituents making up and determining a type of phrase. On the other hand, Johnson (1998) suggests a sort of information propagation in a top-down fashion—a sort of after the fact description of a phrase's context within a given tree. Second, we are not carrying out transformations to be undone after some parsing process; we are carrying out static annotation of treebank trees.

(3) [...] qui risquait de brouiller
[...] who was risking of shake-up
l'image [...]
the image [...]
'who risked messing up the image'[5]

3.4 Raised Coordination

Coordination in the P7T is represented as a sort of adjunction of a COORD phrase as a sister or daughter of the element it is to be coordinated with. This is interpreted in two different ways in the treebank, illustrated in Figure 4, making coordinated structures in the P7T highly ambiguous and inconsistent. Either of the two analyses shown are attested in the P7T, as well as a third, sometimes, for PP coordination.[6,7]

The coordination analysis adopted for the MFT is similar to that of the Penn Treebank (Bies et al. (1995)), except for one important fact: we do not get rid of the COORD constituent. Coordination has been modified to be structured as a single phrase consisting of coordinate daughters. This process of restructuring was carried out in a semi-automatic fashion. All sentences had to be hand corrected after automatic transformation, due to the ambiguity in the structures of the P7T. Generally, the goal of the transformation was to arrive at a structure such as the one in Figure 5, from those in Figure 4 (as well as from any other erroneous coordinated structures encountered).

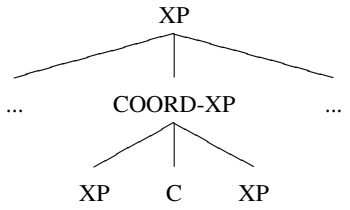

Figure 5: MFT coordination with arguments.

For like-constituent coordination, COORD XML elements now have a "type" attribute, whose value is the type of coordinated constituent (i.e., NP, AP, etc.). In Figure 5, the COORD phrase is of type XP. In addition, it is enclosed in an XP phrase along with any of its shared arguments or modifiers.

Nonconstituent coordination and unlike constituent coordination required slightly different, but similarly structured, analyses. Unlike constituent coordination was labeled with the type UC, and nonconstituent coordination with the type NC, or VP in the case of an NC that really corresponds to a VP.[8]

COORD-UC phrases may take a functional label if they are sister to a VN, whereas COORD-NC phrases do not. In NC coordination, parallel elements are enclosed in a special NC phrase, if they are not argumentally complete verbal phrases phrases (for example, argument cluster coordination). The functional roles of each of their constituents is given on the constituents themselves within the NC or Sint constituent.[9] Figure 6 illustrates a type of NC coordination for the following example.

(4) la personalité morale de la Cinq
the personality moral of the Five
disparaît, et avec elle l'autorisation
disappears, and with her the authorisation
d'émettre
of broadcast
'the moral personality of the Five is disappearing, and with it the permission to broadcast' [10]

3.5 Functional Path Tags

Approximately half of the P7T was automatically functionally annotated and hand corrected (Abeillé and Barrier (2004), cf. Section 2). In the original subsection of the P7T (before being modified and hand corrected by the present authors) the functional tag counts are as given in Table 3.

The functional annotation scheme adopted for the P7T assumed that all sisters of the VN phrase are functionally dependent on that phrase. However, this is not always the case; it-cleft construc-

[5] Sentence 8009, file flmf3_08000_08499ep.xd.cat.xml.

[6] The annotation guidelines of the P7T suggest that there is a difference in distribution, however, upon working with the P7T, one realises that this is not the case. It seems that the flatness of analyses in the trees of the P7T combined with their analysis of coordination has resulted in confused structures. Thus, for any type of constituent coordination, we possibly find either of the different structures in the P7T indicated in Figure 4.

[7] We have also found another regularly used form of coordination for PP coordination, where a PP is coordinated with the mother node of its mother node. However, we think that this is perhaps a consistent error, and not an analysis.

[8] Recall that VP is not a constituent in the P7T, and is not introduced into the MFT, except where NC would correspond to a VP.

[9] In reality, like VN, NC is not really a phrase; rather, it is a convention permitting the expression of parallel structures. We use explicity the tag "NC" to make this clear.

[10] Sentence 154, file flmfaa1ep.cat.xml.

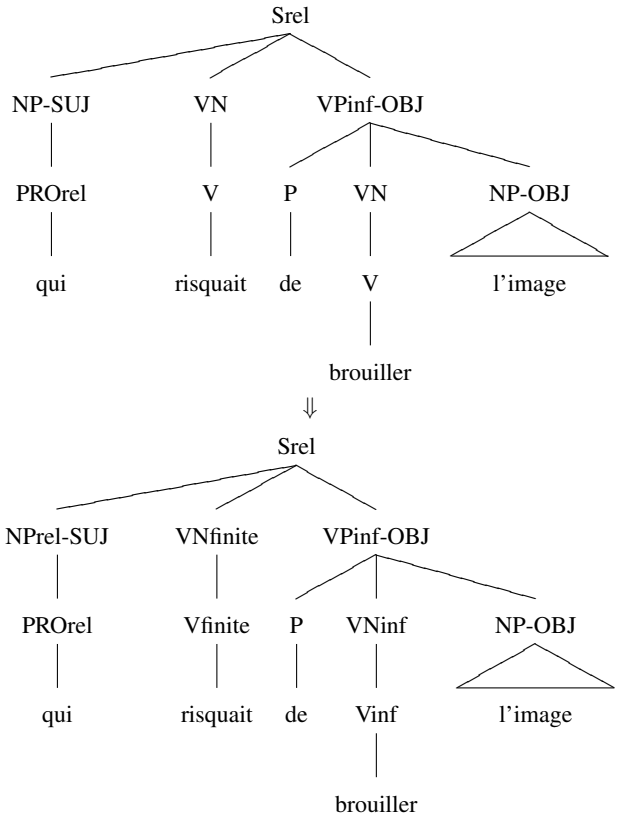

Figure 3: P7T (above) and MFT representation (below) of example (3).

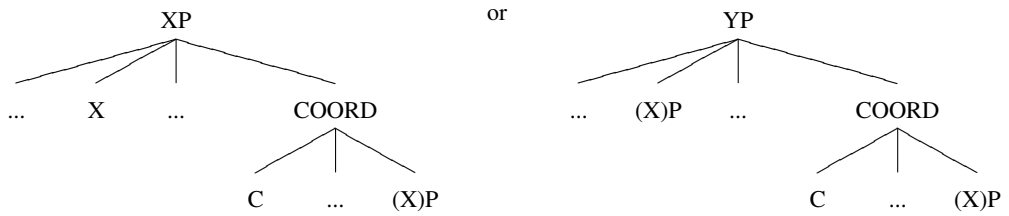

Figure 4: Coordination with Mother or with Sister Node in the P7T. The (X)P are coordinated.

Functional Tag	count	Functional Tag	count
SUJ	8036	OBJ	5949
MOD	6023	A-OBJ	833
DE-OBJ	1354	P-OBJ	913
ATS	560	ATO	104

Table 3: Original Functional Tag Counts for the Relevant P7T Subset

tions provide a first example (cf. Section 3.2). Other cases involve, for example, pronouns for DE prepositional phrases (pronouns such as *dont* or *en*) and daughters of NC. Inspired by the functional paths in the LFG framework[11], we assign new path functions, as illustrated in Figure (2), where the Srel constituent takes the functional path tag SUJ.MOD, representing the fact that Srel has the function MOD, and is dependent on the constituent whose function is SUJ.

We note, in addition, that much of the functional annotation was missing in the functionally annotated subset of the P7T; only 23,772 functional tags were found in the relevant subsection of the P7T. In contrast, the MFT contains 30,399 functional tags. Table 4 presents the MFT counts of the new functional path tags.

4 Formatting and Error Mining

In order to be usable by software, and before any restructuring of the P7T could take place, we carried out an extensive clean-up of the original P7T

[11] See, for example, Dalrymple (2001).

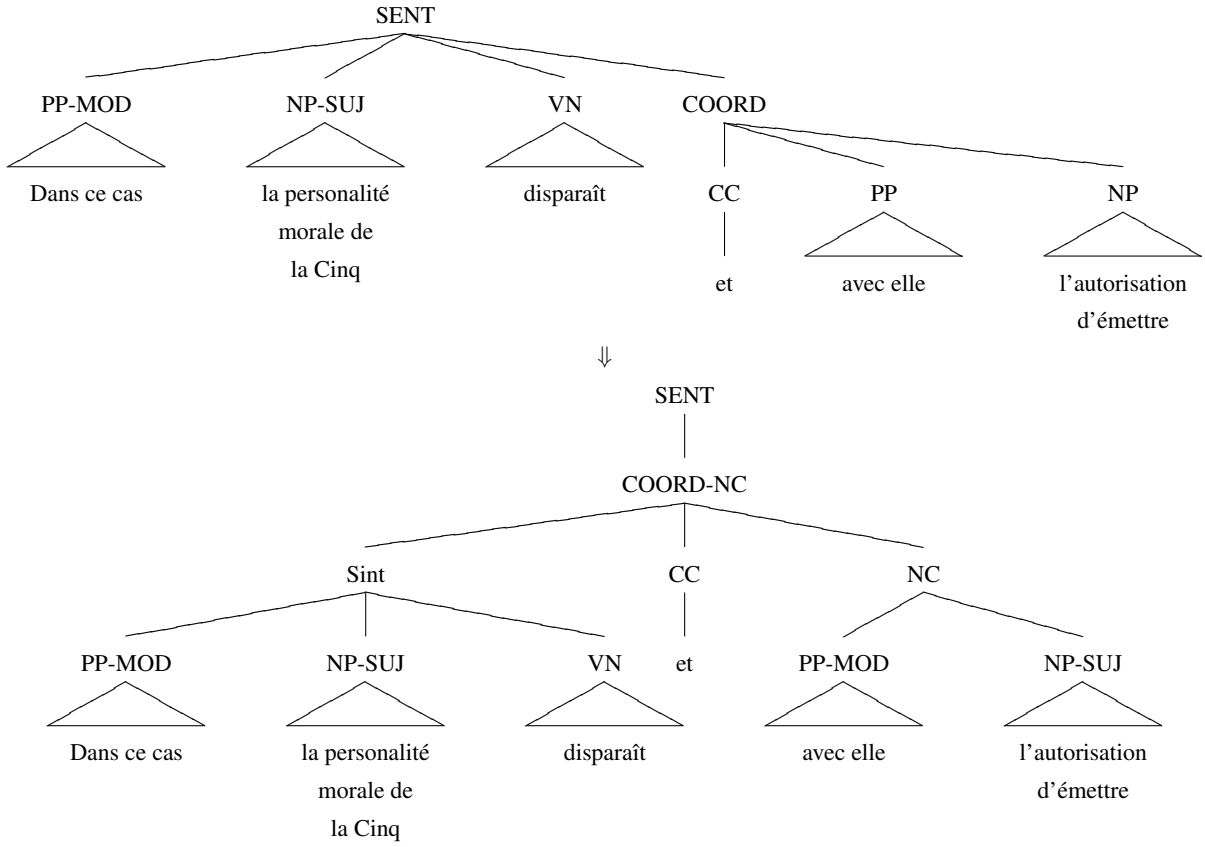

Figure 6: P7T (above) and MFT (below) representation of example (4).

Functional Tag	Count
SUJ	7969
OBJ	6667
MOD	10615
A-OBJ	1432
DE-OBJ	956
ATS	1470
SUJ.MOD	158
P-OBJ	1022
ATO	126
A-OBJ.OBJ	1
ATS.MOD	14
DE-OBJ.OBJ	1
OBJ.MOD	38
OBJ.DE-OBJ	1
OBJ.OBJ	3
SUJ.A-OBJ	1
DE-OBJ.OBJ.MOD	2
OBJ.A-OBJ	2
SUJ.DE-OBJ	1
A-OBJ.OBJ.MOD	1

Table 4: MFT Counts of Functional Path Tags

formatting. This involved, for example, reinserting missing part-of-speech tags, and repairing the XML formatting.[12]

[12]For example, in the whole of the functionally annotated section of the P7T, we found 5 empty SENT constituents, 3 cases of word-forms floating outside of their XML elements, 15 misformatted lemmas, 24 missing parts-of-speech for words not belonging to a multi-word expression, 16,222 missing parts-of-speech for words belonging to a multi-word expression, 18 misused attributes, etc.

Following the reformatting and restructuring of the treebank, a phase of general error mining and correction was undertaken to reduce any noise that we had introduced into the new MFT version of the treebank, and to try to catch any important errors that we had as yet left untreated or that we had missed. Error mining has been shown to improve the results of even very robust techniques for comparatively large corpora (Dickinson and Meurers (2005, 2003a,b)).

This phase has been carried out semi-automatically, in three steps. The first step simply involved automatically extracting a CFG grammar from the treebank, and verifying manually that the productions were consistent with P7T and MFT annotation guidelines, correcting any deviations. The next two steps consisted of applying error-mining software created under the Decca project (Dickinson and Meurers (2005)). This involved applying software for the detection

of part-of-speech variations and constituent-string relation variations, examining non-fringe results, and manually correcting any detected erroneous annotations.[13]

5 Comparative Statistics

The comparative counts of tokens and types of CFG rules for the relevant subset of sentences, given a certain left-hand side, is presented in Table 5.[14] We observe that in all instances (except for AdP[15]), the number of tokens has increased from P7T to MFT, whereas, except for Sint, the number of types has decreased. In addition, we have used a 2-sample χ^2 test for equality to show that all type-token proportions have significantly decreased, as shown in the last column of Table 5 (ranging from the largest P-value 2.546E-02 to the smallest 2.2E-16). The differences reflect the consistency and comparative simplicity of the MFT with respect to the P7T.

Left Side	P7T types/tokens	MFT types/tokens	p-value
SENT	1476/4741	1114/4739	2.2E-16
AP	93/5506	64/8440	5.428E-07
AdP	37/290	44/5755	2.2E-16
NP (or NPrel)	1086/34747	690/38036	2.2E-16
PP (or PPrel)	129/19071	60/19930	1.416E-07
VPinf	300/2940	221/3047	6.229E-05
VPpart	249/2009	160/2115	2.838E-07
Srel	302/1567	233/1590	6.475E-04
Ssub	361/1426	284/1513	2.235E-05
Sint	191/597	273/1024	2.546E-02

Table 5: Productions of the P7T versus the MFT

6 Parsing Results and Regression Analysis

Arun and Keller (2005) report parsing results on the P7T. Post-publication, Arun discovered (personal communication) that the results reported in these publications were erroneously obtained; Arun and Keller (2005) mistakenly discarded over half of the treebank trees, believing that the contracted words were XML errors. Their new results for sentences of length ≤ 40 words were given in their presentation at ACL, and are reported in Table 6.[16]

Parser and Mode	LR	LP	f-score
BitPar (own POS tagging)	64.49	64.36	64.42
BitPar (perfect tagging mode)	67.78	67.07	67.42
Bikel (own POS tagging)	79.94	79.36	79.65
Bikel (unknown POS supplied)	80.79	80.23	80.50

Table 6: Arun and Keller's P7T parsing results ($\leq$ 40 words)

Arun and Keller present results for BitPar (Schmid (2004)), as well as for several modifications made to Bikel's parser (Bikel (2002)). What they term as "Collins Model 2" is essentially Bikel's parser without any of the added modifications; results from this model applied to the best of the Arun and Keller (2005)'s transformations of the P7T (contracted compounds and raised coordination[17]) will serve as a baseline for comparison with our results here.

Upon finding that Bikel's parser outperforms BitPar when trained on the P7T by over 15%, Arun and Keller concluded that French, like English but unlike German, parses best in a lexicalised statistical parsing framework, leading to the conjecture that word order, and not flatness of annotation, is crucial for lexicalisation. By contrast, parsing results with the MFT lead to a less extreme conclusion, and provide further evidence that a coherent and well-structured treebank leads to better parsing results.

Experiments were repeated on the MFT using both BitPar and Bikel's parser. The MFT was randomly subdivided into a training set (3800 sentences), development set (509 sentences) and a test set (430 sentences). Our training set roughly corresponds (in quantity) to only 20.5% of the training data used by Arun and Keller in their most recent experiments (18,548 sentences), yet our re-

[13] For example, Decca POS software detects 28 7-gram variations of which 15 are non-fringe. The non-fringe variations were examined for errors. The same softwares detects only 11 7-gram variations in the MFT, of which 5 are non-fringe.

[14] COORD and VN, and any new constituents added to the MFT are not mentioned for reasons of incomparability. Also note that these rule counts make abstraction of any punctuation or functional tagging.

[15] Observe that the AdP phrase in the original P7T was comparatively rarely employed.

[16] The ACL slides presenting these new results may be obtained at http://homepages.inf.ed.ac.uk/s0343799/acl2005slides.pdf .

[17] As with Arun and Keller's work, we contracted compounds for our experiments. Their method of raising coordination is completely different from the one discussed here. See (Arun and Keller (2005)) for details.

sults show improvements on results using the P7T. The results for the MFT are shown in Table 7.

Parser and Mode	LR	LP	f-score
BitPar (own POS tagging)	70.66	70.62	70.64
BitPar (perfect tagging mode)	78.07	77.36	77.71
Bikel (own POS tagging)	79.76	80.13	79.95
Bikel (unknown POS supplied)	83.09	83.31	83.20
Bikel (perfect tagging mode)	84.62	84.69	84.66

Table 7: MFT parsing results ($\leq$ 40 words)

BitPar trained on the MFT outscores across the board its scores when trained on more than five times the amount of data from the P7T. On sentences of length less than 40 words, BitPar trained on the MFT scores 6.22% better, and in perfect tagging mode, BitPar scores 10.29% better than when trained on the substantially larger training set from the P7T.

Smaller increases are also achieved for Bikel's parser, when trained on the small training set of the MFT. When Bikel's parser carries out its own POS tagging, it scores 0.3% better, and when unknown POS tags are supplied, it performs 2.51% better than its counterpart trained on the large training set of the P7T.

Table 7 also shows how scores using Bikel's parser jump, once again, when run in perfect tagging mode. Arun and Keller do not report results for running Bikel in perfect tagging mode.[18]

The variances in the increases of f-scores seem to be the direct results of the parsing mechanisms adopted by each of the parsers. BitPar is less flexible to inconsistent and error-ridden data, than Bikel's parser, which assumes independence relations among sister nodes, compensating for this with only a distance measurement.

The learning curves in Figure 7 present the changes in parser performance trained on increasingly large subsets of the MFT training set. For this experiment, we also train on the development set to obtain further information about possible increases in parser performance and its possible simple correlation to training set size.

Due to the small number of observations, any nonlinear growth curve fitting method would be parsimonious, we therefore applied linear regression analysis. Using four different combinations of power transformations, we found these learning curves to be approximately linear with a very strong positive relationship between transformed number and f-score. Table 8 shows the transforms, R^2, parameters, and parameter p-values (using the standard t-test). F-score extrapolation for a training set of size 18,548 (the size of the training set for experiments by Arun and Keller on the P7T) are given in Table 9. These predictions show an increase in f-score across the board.[19]

Parser and Mode	P7T f-score	MFT predicted f-score
BitPar (own POS tagging)	64.42	75.72
BitPar (perfect tagging)	67.42	81.08
Bikel (own POS tagging)	79.65	82.44
Bikel (unknown POS supplied)	80.50	83.99

Table 9: F-score and f-score prediction comparison for training set of size 18,548

The largest increase between MFT parsing scores and predicted parsing scores with a larger training set is for BitPar, whose predicted score is 11.3% higher when doing its own POS tagging, and 13.66% higher in perfect tagging mode. In fact, the performance gap between BitPar and Bikel's parser seems to be steadily closing as MFT training data sizes increase. These results suggest that lexicalisation for statistical parsing of French is perhaps not as crucial as was concluded by Arun and Keller (2005).[20]

7 Conclusion

We have presented the Modified French Treebank, a new French Treebank, derived from the P7T, which is cleaner, more coherent, has several transformed structures, and introduces new linguistic analyses. The positive effect of transformations on and cleaning up treebanks is well documented (for example, by Dickinson and Meurers (2005)). We investigated one important effect of a clean treebank on corpus linguistics. The MFT provides an

[18] Bikel's parser can be tricked into perfect tagging mode, by appending the part-of-speech to the end of each word-form.

[19] Significance tests are not applicable.

[20] Some authors (for example, Rehbein and van Genabith (2007); Kűbler (2005)) argue that parsing results for treebanks with different annotation schemes are not comparable. However, this conclusion remains unaffected by such argumentation.

Parser and Mode	transform	R^2	α	p-value (α)	β	p-value (β)
BitPar (own POS tagging)	$y = \alpha \cdot ln(x) + \beta$	0.9978	6.7206	1.53E-10	14.8734	8.11E-07
BitPar (perfect tagging)	$y = \frac{1}{\alpha \cdot ln(x) + \beta}$	0.9616	-0.0003	3.283E-06	0.0156	1.1472E-11
Bikel (own POS tagging)	$y = \frac{\alpha}{ln(x)} + \beta$	0.9943	-298.6927	4.03E-09	115.4334	2.42E-12
Bikel (unknown POS supplied)	$y = \frac{ln(x)}{\alpha + \beta \cdot ln(x)}$	0.977	0.01551	5.42E-07	0.0102	8.32E-12

Table 8: Linear regression on learning curve data from Figure 7

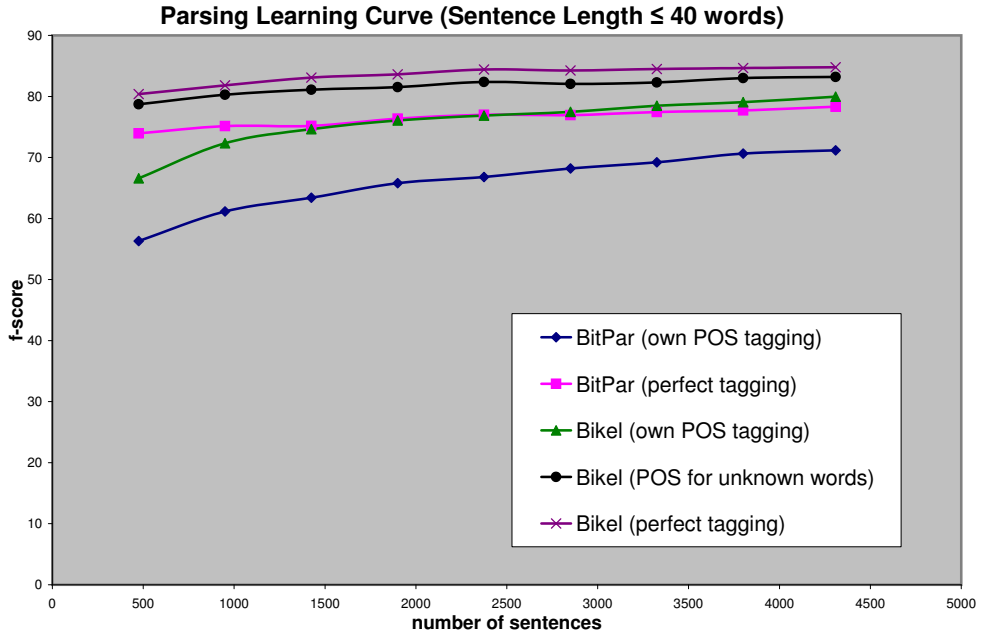

Figure 7: Learning Curve for the MFT.

extreme example of how quantity does not always make up for quality in statistical parsing. A probabilistic parser trained on clean and transformed data performs better than its counterpart trained on the original French treebank, which consists of five times the data. Moreover, we have shown how data which has a high error rate and that is not "parser-friendly" can lead to the potentially erroneous conclusions about the impact of lexicalisation on probabilistic parsing of French.

Acknowledgements

We gratefully acknowledge support from Science Foundation Ireland grant 04/IN/I527 for the research reported in this paper.

References

Anne Abeillé and Nicholas Barrier. 2004. Enriching a french treebank. In *LREC Conference Proceedings*. Lisbon.

Anne Abeillé, François Toussenel, and Martine Chéradame. 2004. Corpus le monde: Annotations en constituants. guide pour les correcteurs. Technical report, LLF and UFRL and Université Paris 7.

A. Arun and F. Keller. 2005. Lexicalisation in crosslinguistic probablisitic parsing: The case of french. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 306–313. Ann Arbor, MI.

A. Bies, M. Ferguson, K. Katz, and R. MacIntyre. 1995. Bracketing guidelines for treebank ii style penn treebank project. Technical report, University of Pennsylvania.

D. Bikel. 2002. Design of a multi-lingual, parallel-processing statistical parsing engine. In *Proceedings of the 2nd International Conference on Human Language Technology Research*. San Francisco.

Mary Dalrymple. 2001. *Lexical Functional Grammar*, volume 34 of *Syntax and Semantics*. Academic Press, San Diego.

Markus Dickinson and W. Detmar Meurers. 2003a. Detecting errors in part-of-speech annotation. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 107–114. Budapest, Hungary.

Markus Dickinson and W. Detmar Meurers. 2003b. Detecting inconsistencies in treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, pages 45–56. Växjö, Sweden.

Markus Dickinson and W. Detmar Meurers. 2005. Prune diseased branches to get healthy trees! how to find erroneous local trees in a treebank and why it matters. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*. Barcelona, Spain.

Mark Johnson. 1998. Pcfg models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.

S. Kűbler. 2005. How do treebank annotation schemes influence parsing results? or how not to compare apples and oranges. In *Proc. of RANLP*. Borovets, Bularia.

I. Rehbein and J. van Genabith. 2007. Treebank annotation schemes and parser evaluation for german. In *Proc. of the EMNLP-CoNLL 2007*. Prague, Czech Republic.

H. Schmid. 2004. Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland.

Leonoor van der Beek. 2003. The dutch it-cleft constructions. In *Proceedings of the LFG03 Conference*.